

# Asymptotic power of likelihood ratio tests for high dimensional data

Cheng Wang<sup>\*a,b</sup>, Longbing Cao<sup>†b</sup> and Baiqi Miao<sup>‡a</sup>

<sup>a</sup>Department of Statistics and Finance, University of Science and Technology of China, Hefei, Anhui 230026, China

<sup>b</sup>Advanced Analytics Institute, University of Technology Sydney, New South Wales 2007, Australia

## Abstract

This paper considers the asymptotic power of likelihood ratio test (LRT) for the identity test when the dimension  $p$  is large compared to the sample size  $n$ . The asymptotic distribution of LRT under alternatives is given and an explicit expression of the power is derived when  $\text{tr}(\Sigma_p) - \log |\Sigma_p| - p$  tends to a constant. A simulation study is carried out to compare LRT with other tests. All these studies show that LRT is a powerful test to detect eigenvalues around zero.

Key words and phrases: Covariance matrix, High dimensional data, Identity test, Likelihood ratio test, Power

## 1 Introduction

In multivariate analysis for high dimensional data, testing the structure of population covariance matrices is an important problem. See, for example, Johnstone (2001), Ledoit and Wolf (2002), Srivastava (2005), Schott (2006), Chen *et al.* (2010), Cai and Jiang (2011) and Li and Chen (2012), among many others. Let  $X_1, \dots, X_n$  be  $n$  independent and identically distributed (i.i.d.)  $p$ -variate random vectors following a multivariate normal distribution  $N_p(\mu, \Sigma_p)$  where  $\mu$  is the mean vector and  $\Sigma_p$  is the covariance matrix. In many studies, a hypothesis test of significant interest is to test

$$H_0 : \Sigma_p = I_p \text{ vs } H_1 : \Sigma_p \neq I_p, \quad (1)$$

---

<sup>\*</sup>wwcc@mail.ustc.edu.cn

<sup>†</sup>LongBing.Cao@uts.edu.au

<sup>‡</sup>bqmiao@ustc.edu.cn

where  $I_p$  is the  $p$ -dimensional identity matrix. Note that the identity matrix in (1) can be replaced by any other positive definite matrix  $\Sigma_0$  through multiplying the data by  $\Sigma_0^{-1/2}$ .

A natural approach to test (1) is to conduct estimations for some distance measures between  $\Sigma_p$  and  $I_p$  and there are two types of measures which are widely used in literature. The first is based on the likelihood function:

$$L_l(\Sigma_p) = \text{tr}(\Sigma_p) - \log |\Sigma_p| - p, \quad (2)$$

and the second is based on quadratic loss function:

$$L_q(\Sigma_p) = \text{tr}(\Sigma_p - I_p)^2. \quad (3)$$

Each of these is 0 when  $\Sigma_p = I_p$  and positive when  $\Sigma_p \neq I_p$ . For  $L_l(\Sigma_p)$ , it is referred to the likelihood ratio test (LRT) and the classic LRT for fixed  $p$  and large  $n$  can be found in Anderson (2003). For high dimensional data ( $p$  is large), the failure of classical LRT was firstly observed by Dempster (1958) and later in a pioneer work by Bai *et al.* (2009), authors proposed corrections to LRT when  $p/n \rightarrow c \in (0, 1)$  and  $\mu = 0$ . Successive works included Jiang *et al.* (2012) which extended the results of Bai *et al.* (2009) to Gaussian data with general  $\mu$  and our work Wang *et al.* (2012) where we studied the LRT for general  $\mu$  and non-Gaussian data. For the quadratic loss function  $L_q(\Sigma_p)$ , there are many works since the seminal paper Nagao (1973). These included Ledoit and Wolf (2002), Birke and Dette (2005), Srivastava (2005), Chen *et al.* (2010) and Cai and Ma (2012). Other works which considered question (1) are referred to Johnstone (2001), Cai and Jiang (2011) and Onatski *et al.* (2011).

The existing results about LRT (Bai *et al.*, 2009; Jiang *et al.*, 2012; Wang *et al.*, 2012) have only derived asymptotic null distribution and we know little about the asymptotic point-wise power of LRT under the alternative hypothesis. Recently, Onatski *et al.* (2011) studied the asymptotic power of several tests including LRT under the special alternative of rank one perturbation to the identity matrix as both  $p$  and  $n$  go to infinity. Cai and Ma (2012) investigated the testing problem (1) in the high-dimensional settings from a minimax point of view. The results in Onatski *et al.* (2011) and Cai and Ma (2012) showed that LRT was a sub-optimal test when there was a rank one perturbation to the identity matrix.

In this work, we will consider the power of LRT under general alternatives. The asymptotic distribution of LRT will be studied when  $\Sigma_p \neq I_p$  and an explicit expression of the power will also be derived when  $L_l(\Sigma_p)$  tends to a constant. From these results, we find that in relation to LRT it is not fair that Onatski *et al.* (2011) and Cai and Ma (2012) only focused on the alternatives whose eigenvalues were larger than 1. Furthermore, our results show that LRT is powerful to detect eigenvalues around zero. Simulations will also be conducted to compare LRT with two tests based on  $L_q(\Sigma_p)$  (Chen *et al.*, 2010; Cai and Ma, 2012).

The paper is structured as follows: Section 2 introduces the basic data structure and establishes the asymptotic power of LRT while Section 3 reports simulation studies. All the proofs are presented in the Appendix.

## 2 Main Results

To relax the Gaussian assumptions, we assume that the observations  $X_1, \dots, X_n$  satisfy a multivariate model (Chen *et al.*, 2010)

$$X_i = \Sigma_p^{1/2} Y_i + \mu, \text{ for } i = 1, \dots, n \quad (4)$$

where  $\mu$  is a  $p$ -dimensional constant vector and the entries of  $\mathcal{Y}_n = (Y_{ij})_{p \times n} = (Y_1, \dots, Y_n)$  are i.i.d. with  $EY_{ij} = 0$ ,  $EY_{ij}^2 = 1$  and  $EY_{ij}^4 = 3 + \Delta$ . The sample covariance matrix is defined as

$$S_n = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})(X_k - \bar{X})',$$

where  $\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k$ .

Writing  $y_n = p/n < 1$ , the LRT statistic is defined as

$$L_n = \frac{1}{p} \text{tr}(S_n) - \frac{1}{p} \log |S_n| - 1 - d(y_n) \quad (5)$$

where  $\text{tr}$  denotes the trace and  $d(x) = 1 + (1/x - 1) \log(1 - x)$ ,  $0 < x < 1$ . Under the null hypothesis, Wang *et al.* (2012) derived the following asymptotic normality of  $L_n$  by using random matrix theories.

**Theorem 1 (Theorem 2.1 of Wang *et al.* (2012))** When  $\Sigma_p = I_p$  and  $y_n = p/y \rightarrow y \in (0, 1)$ ,

$$\frac{pL_n - \mu_n}{\sigma_n} \xrightarrow{D} N(0, 1),$$

where  $\mu_n = y_n(\Delta/2 - 1) - 3/2 \log(1 - y_n)$ ,  $\sigma_n^2 = -2y_n - 2 \log(1 - y_n)$  and  $\xrightarrow{D}$  denotes convergence in distribution.

When  $X_1, \dots, X_n$  be i.i.d. multivariate normal distributions  $N_p(\mu, \Sigma_p)$  where  $\Delta = 0$ , Jiang *et al.* (2012) derived a similar result as Theorem 1 by using the Selberg integral and they also considered the special situation where  $p/n \rightarrow 1$ . Based on the asymptotic normality under the respective null hypothesis, an asymptotic level  $\alpha$  test based on  $L_n$  is given by

$$\phi = I\left(\frac{pL_n - \mu_n}{\sigma_n} > z_{1-\alpha}\right), \quad (6)$$

where  $I(\cdot)$  is the indicator function, and  $z_{1-\alpha}$  denotes the  $100 \times (1 - \alpha)$ th percentile of the standard normal distribution. In the following theorem, we establish the convergence of  $L_n$  under the alternative  $\Sigma_p \neq I_p$ .

**Theorem 2** When  $\text{tr}(\Sigma_p - I_p)^2/p \rightarrow 0$  and  $y_n = p/y \rightarrow y \in (0, 1)$ ,

$$\frac{pL_n - L_l(\Sigma_p) - \mu_n}{\sigma_n} \xrightarrow{D} N(0, 1),$$

where  $\mu_n = y_n(\Delta/2 - 1) - 3 \log(1 - y_n)/2$  and  $\sigma_n^2 = -2y_n - 2 \log(1 - y_n)$ .

In particular, when  $L_l(\Sigma_p)$  tends to a constant, we have the following result.

**Theorem 3** When  $L_l(\Sigma_p) \rightarrow b \in (0, \infty)$  and  $y_n = p/y \rightarrow y \in (0, 1)$ ,

$$\lim_{n \rightarrow \infty} P_{\Sigma_p}(\phi \text{ rejects } H_0) = 1 - \Phi\left(z_{1-\alpha} - \frac{b}{\sqrt{-2y - 2\log(1-y)}}\right),$$

where  $\Phi(\cdot)$  is the cumulative distribution function of the standard normal distribution.

It can be seen from Theorems 2 and 3 that the expression

$$1 - \Phi\left(z_{1-\alpha} - \frac{L_l(\Sigma_p)}{\sigma_n}\right), \quad (7)$$

gives good approximation to the power of the test in (6) until the power is extremely close to 1. In particular, when  $L_l(\Sigma_p)$  is large, the power of the test  $\phi$  will be close to 1 and it is hard for  $\phi$  to distinguish between the two hypotheses if  $L_l(\Sigma_p)$  tends to zero.

To derive the asymptotic power, a special covariance matrix was used in Cai and Ma (2012) and Onatski *et al.* (2011) as follows

$$\Sigma_p^* = I_p + h\sqrt{\frac{p}{n}}vv', \quad (8)$$

where  $h$  is a constant and  $v$  is an arbitrarily fixed unit vector. For this special spiked matrix (Johnstone, 2001), the true  $\Sigma_p$  has a perturbation in a single unknown direction and in Cai and Ma (2012) and Onatski *et al.* (2011), the authors focused on situations where  $h > 0$  that is one eigenvalue of  $\Sigma_p$  is larger than 1 while others are still unitary. Here, by Theorem 3, we know that the asymptotic power of the LRT test (5) is

$$1 - \Phi\left(z_{1-\alpha} - \frac{h\sqrt{y} - \log(1 + h\sqrt{y})}{\sqrt{-2y - 2\log(1-y)}}\right).$$

Therefore, compared with the tests based on  $L_q(\Sigma_p)$  (Ledoit and Wolf, 2002; Chen *et al.*, 2010; Cai and Ma, 2012) whose power for  $\Sigma_p^*$  is  $1 - \Phi(z_{1-\alpha} - h^2/2)$ , LRT is more sensitive to small eigenvalues ( $h < 0$ ), not any bigger than one ( $h > 0$ ). In particular, when  $1 + h\sqrt{y}$  is close to 0 that is  $\Sigma_p^*$  has a very small eigenvalue, the power will tend to 1. A numerical experiment will be conducted in the next section to show the performances of LRT (6) and the tests based on  $L_q(\Sigma_p)$ .

### 3 Simulations

In this section, we conduct several simulation studies to compare the power of the LRT in (6) with that of the tests based on  $L_q(\Sigma_p)$ . When  $X_1, \dots, X_n$

i.i.d from  $N_p(0, \Sigma_p)$ , Cai and Ma (2012) proposed an estimator of  $L_q(\Sigma_p)$  as

$$T_{1,n} = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} h(X_i, X_j), \quad (9)$$

where  $h(X_1, X_2) = (X_1' X_2)^2 - (X_1' X_1 + X_2' X_2) + p$  and an asymptotic level  $\alpha$  test based on  $T_{1,n}$  is given by

$$\phi_1 = I(T_{1,n} > z_{1-\alpha} 2 \sqrt{\frac{p(p+1)}{n(n-1)}}). \quad (10)$$

Similarly, for general data structure, Chen *et al.* (2010) gave an estimator for  $L_q(\Sigma_p)$  as

$$\begin{aligned} T_{2,n} = & \frac{1}{P_n^2} \sum_{i,j}^* (X_i' X_j)^2 - \frac{2}{P_n^3} \sum_{i,j,k}^* X_i' X_j X_j' X_k \\ & + \frac{1}{P_n^4} \sum_{i,j,k,l}^* X_i' X_j X_k' X_l - 2tr(S_n) + p, \end{aligned}$$

where  $P_n^r = n!/(n-r)!$  and  $\sum^*$  denotes summation over mutually different indices. And the level  $\alpha$  test is

$$\phi_2 = I(\frac{n}{2p} T_{2,n} > z_{1-\alpha}). \quad (11)$$

To evaluate the power of the tests, the alternative population covariance matrix will be set as  $\Sigma^* = \text{diag}(\rho, 1, \dots, 1)$  where  $\rho$  will range from 0.01 to 4. Here, we only focus on this simple matrix to investigate the features of each test and simulations for more general alternatives can be found in our previous work Wang *et al.* (2012). All the results are based on  $10^4$  replications.

Figure 1 reports empirical powers of the LRT and the test of Cai and Ma (2012) (CM test) for  $N_p(0, \Sigma_p)$  distributions. We observe from Figure 1 that LRT had a better performance for  $\rho < 1$  while the CM test performed better if  $\rho > 1$ . When  $\rho$  is around 1 that is the true  $\Sigma_p$  is very close to an identity matrix, both tests had similar empirical sizes which were quite close to the nominal 5%. For the fixed sample size  $n = 200$ , when  $p$  was increased from 50 to 100, the performances of both tests became poor which could be understood as the estimators get worse as  $p$  is increased for the fixed sample size.

For general data, due to the CM test is only applicable for  $N_p(0, \Sigma_p)$ , we conduct comparisons between LRT and the test of Chen *et al.* (2010) (CZZ test) where  $Y_{ij}$  comes from Gamma(4, 0.5) distribution in data (4). Figure 2 reports the empirical powers of the LRT and CZZ tests and the conclusions follow very similar patterns to those of Figure 1 which shows that LRT is a powerful test to detect eigenvalues around zero.

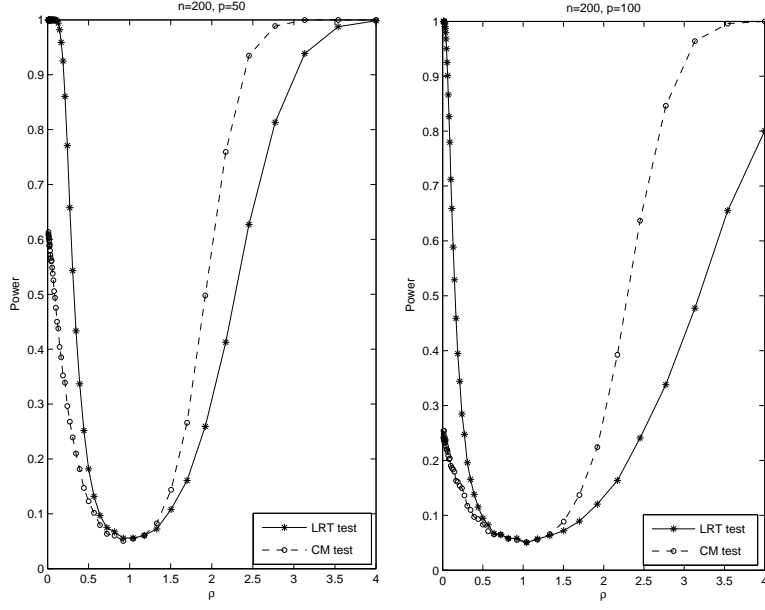


Figure 1: Empirical performances of LRT and CM tests for Gaussian data with known means.

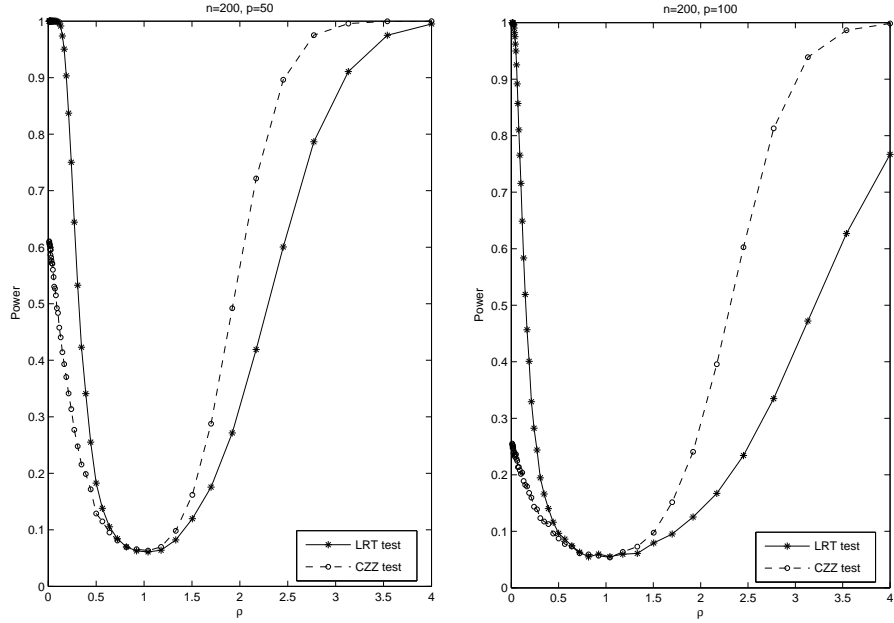


Figure 2: Empirical performances of LRT and CZZ tests for non-Gaussian data with unknown means

## 4 Appendix

### 4.1 Proof of Theorem 2

Writing

$$B_n = \frac{1}{n-1} \sum_{k=1}^n (Y_k - \bar{Y})(Y_k - \bar{Y})', \quad (12)$$

where  $\bar{Y} = \sum_{k=1}^n Y_k$ . Noting  $S_n = \Sigma_p^{1/2} B_n \Sigma_p^{1/2}$ , we have

$$\begin{aligned} L_n &= \frac{1}{p} \text{tr}(S_n) - \frac{1}{p} \log |S_n| - 1 - d(y_n) \\ &= \frac{1}{p} \text{tr}(\Sigma_p B_n) - \frac{1}{p} \text{tr}(B_n) - \frac{1}{p} \log |\Sigma_p| + BL_n \\ &= BL_n + \frac{1}{p} L_l(\Sigma_p) + \frac{1}{p} \text{tr}((\Sigma_p - I_p) B_n) - \frac{1}{p} \text{tr}(\Sigma_p - I_p), \end{aligned}$$

where  $BL_n = \frac{1}{p} \text{tr}(B_n) - \frac{1}{p} \log |B_n| - 1 - d(y_n)$ . By Theorem 1,

$$\frac{pBL_n - \mu_n}{\sigma_n} \xrightarrow{D} N(0, 1).$$

Therefore, to prove Theorem 2, it is enough to show

$$\epsilon_n := \text{tr}((\Sigma_p - I_p) B_n) - \text{tr}(\Sigma_p - I_p) = o_p(1).$$

Rewriting  $B_n$  as

$$B_n = \frac{1}{n} \sum_{k=1}^n Y_k Y_k' - \frac{1}{n(n-1)} \sum_{i \neq j} Y_i Y_j', \quad (13)$$

we can get  $E[\epsilon_n] = 0$  and by Proposition A.1 of Chen *et al.* (2010),

$$\begin{aligned} E[\epsilon_n^2] &= E\left[\left(\frac{1}{n} \sum_{k=1}^n Y_k'(\Sigma_p - I_p) Y_k - \frac{1}{n(n-1)} \sum_{i \neq j} Y_i'(\Sigma_p - I_p) Y_j\right)^2\right] - (\text{tr}(\Sigma_p - I_p))^2 \\ &= \frac{2}{n-1} \text{tr}((\Sigma_p - I_p)^2) + \frac{\Delta}{n} \text{tr}((\Sigma_p - I_p) \circ (\Sigma_p - I_p)) \\ &\leq \frac{2+\Delta}{n-1} \text{tr}((\Sigma_p - I_p)^2) \end{aligned}$$

where  $\circ$  denotes Hadamard product. Above all, when  $p/n \rightarrow y$  and  $\text{tr}((\Sigma_p - I_p)^2)/p \rightarrow 0$ , we come to  $\epsilon_n = o_p(1)$ .

The proof is completed.

## 4.2 Proof of Theorem 3

To prove the theorem, we need some inequalities.

**Lemma 1** *For any  $x > 0$ ,*

- (1) *If  $0 < x \leq 1$ ,  $(x - 1)^2 \leq 2(x - 1 - \log x)$ ;*
- (2) *If  $1 < x < M$ ,  $(x - 1)^2 \leq 2M(x - 1 - \log x)$ .*

Since  $L_l(\Sigma_p) \rightarrow b \in (0, \infty)$ , for large enough  $p$ , we will always have  $D_1(\Sigma_p) < 2b$ . Denoting the eigenvalues of  $\Sigma_p$  as  $\lambda_{1,p}, \dots, \lambda_{p,p}$ ,

$$L_l(\Sigma_p) = \sum_{k=1}^p (\lambda_{k,p} - \log \lambda_{k,p} - 1) < 2b,$$

which implies that there is a constant  $c_0 = c_0(b) > 1$  satisfying

$$c_0^{-1} \leq \lambda_{1,p}, \dots, \lambda_{p,p} \leq c_0.$$

By Lemma 1,  $(\lambda_{k,p} - 1)^2 \leq 2c_0(\lambda_{k,p} - \log \lambda_{k,p} - 1)$  which means

$$L_q(\Sigma_p) = \sum_{k=1}^p (\lambda_{k,p} - 1)^2 \leq 2c_0 L_l(\Sigma_p) < 4c_0 b = o(p).$$

By Theorem 2 and Slutsky's lemma,

$$\frac{pL_n - \mu_n}{\sigma_n} - \frac{b}{\sqrt{-2y - 2\log(1-y)}} \xrightarrow{D} N(0, 1).$$

Now, we can calculate the power of the test

$$\begin{aligned} P_{\Sigma_p}(\phi \text{ rejects } H_0) &= P\left(\frac{pL_n - \mu_n}{\sigma_n} > z_{1-\alpha}\right) \\ &= P\left(\frac{pL_n - \mu_n}{\sigma_n} - \frac{b}{\sqrt{-2y - 2\log(1-y)}} > z_{1-\alpha} - \frac{b}{\sqrt{-2y - 2\log(1-y)}}\right) \\ &\rightarrow 1 - \Phi\left(z_{1-\alpha} - \frac{b}{\sqrt{-2y - 2\log(1-y)}}\right). \end{aligned}$$

The proof is completed.

## Acknowledgement

We thank Dr. Guangming Pan and Dr. Zongming Ma for their helpful discussions and suggestions. The research of Cheng Wang and Baiqi Miao was partly supported by NSF of China Grands No. 11101397 and 71001095. Longbing Cao's research was supported in part by the Australian Research Council Discovery Grant DP1096218 and the Australian Research Council Linkage Grant LP100200774.



## References

- Anderson, T. (2003). *An introduction to multivariate statistical analysis*. Hoboken, NJ:Wiley.
- Bai, Z., Jiang, D., Yao, J., and Zheng, S. (2009). Corrections to LRT on large-dimensional covariance matrix by rmt. *The Annals of Statistics*, **37**(6B), 3822–3840.
- Birke, M. and Dette, H. (2005). A note on testing the covariance matrix for large dimension. *Statistics & Probability Letters*, **74**(3), 281–289.
- Cai, T. and Jiang, T. (2011). Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *The Annals of Statistics*, **39**(3), 1496–1525.
- Cai, T. and Ma, Z. (2012). Optimal hypothesis testing for high dimensional covariance matrices. *arXiv:1205.4219*.
- Chen, S., Zhang, L., and Zhong, P. (2010). Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association*, **105**(490), 810–819.
- Dempster, A. (1958). A high dimensional two sample significance test. *The Annals of Mathematical Statistics*, **29**(4), 995–1010.
- Jiang, D., Jiang, T., and Yang, F. (2012). Likelihood ratio tests for covariance matrices of high-dimensional normal distributions. *Journal of Statistical Planning and Inference*.
- Johnstone, I. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, **29**(2), 295–327.
- Ledoit, O. and Wolf, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *The Annals of Statistics*, pages 1081–1102.
- Li, J. and Chen, S. (2012). Two sample tests for high-dimensional covariance matrices. *The Annals of Statistics*, **40**(2), 908–940.
- Nagao, H. (1973). On some test criteria for covariance matrix. *The Annals of Statistics*, pages 700–709.
- Onatski, A., Moreira, M., and Hallin, M. (2011). Asymptotic power of sphericity tests for high-dimensional data. *Manuscript*.
- Schott, J. (2006). A high-dimensional test for the equality of the smallest eigenvalues of a covariance matrix. *Journal of Multivariate Analysis*, **97**(4), 827–843.
- Srivastava, M. (2005). Some tests concerning the covariance matrix in high dimensional data. *J. Japan Statist. Soc.*, **35**(2), 251–272.

Wang, C., Yang, J., Miao, B., and Cao, L. (2012). On identity tests for high dimensional data using rmt. *arXiv:1203.3278*.